RESEARCH PAPER

The Integration of a Canonical Workflow Framework with an Informatics System for Disease Area Research

Vivek Navale[†] & Matthew McAuliffe

Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892, USA

Keywords: Canonical Workflow Framework for Research; Informatics system; Traumatic brain injury; Parkinson's disease

Citation: Navale, V., McAuliffe, M.: The integration of a canonical workflow framework with an informatics system for disease area research. Data Intelligence 4(2), 186-195 (2022). doi: 10.1162/dint_a_00125
Received: August 4, 2021; Revised: December 7, 2021; Accepted: 1 March, 2022

ABSTRACT

A recurring pattern of access to existing databases, data analyses, formulation of new hypotheses, use of an experimental design, institutional review board approvals, data collection, curation, and storage within trusted digital repositories is observable during clinical research work. The workflows that support the repeated nature of these activities can be ascribed as a Canonical Workflow Framework for Research (CWFR). Disease area clinical research is protocol specific, and during data collection, the electronic case report forms can use Common Data Elements (CDEs) that have precisely defined questions and are associated with the specified value(s) as responses. The CDE-based CWFR is integrated with a biomedical research informatics computing system, which consists of a complete stack of technical layers including the Protocol and Form Research Management System. The unique data dictionaries associated with the CWFR for Traumatic Brain Injury and Parkinson's Disease resulted in the development of the Federal Interagency Traumatic Brain Injury and Parkinson's Disease Biomarker systems. Due to a canonical workflow, these two systems can use similar tools, applications, and service modules to create findable, accessible, interoperable, and reusable Digital Objects. The Digital Objects for Traumatic Brain Injury and Parkinson's disease contain all relevant information needed from the time data is collected, validated, and maintained within a Storage Repository for future access. All Traumatic Brain Injury and Parkinson's Disease studies can be shared as Research Objects that can be produced by aggregating related resources as information packages and is findable on the Internet by using unique identifiers. Overall, the integration of CWFR with an informatics system has resulted in the reuse of software applications for several National Institutes of Health-supported biomedical research programs.

[†] Corresponding author: Vivek Navale (Email: Vivek.Navale@nih.gov; ORCID: 0000-0002-7110-8946).

1. INTRODUCTION

Progress and discoveries in disease area(s) research are possible by promoting the exploration of hypotheses and designing experiments to investigate fundamental biomedical questions that can result in new treatments for patients. A wide range of biomedical platforms with an infrastructure comprising of hardware, software, application, and services can be used to manage research data. The workflows increasingly require the use of high-performance computing for collecting, annotating, curating, analyzing, and storing data within repositories [1].

Despite the diversity and wide-ranging research work on various diseases, a recurring pattern is identifiable. This pattern includes access to existing disease-specific information, analyses of available data, generation of hypotheses, use of an experimental design, review by an institutional review board, and the collection, curation, storage, and preservation of research data within repositories.

A recent white paper has conceptualized the patterns observed within laboratory research as a Canonical Workflow Framework for Research (CWFR) [2]. The elements that constitute a CWFR are the presence of recurring steps in research activities, and the capability to form canonical information packages and libraries. These recurring steps (sequential and/or iterative steps) can result in data and metadata collection, and storage of information packages within repositories.

Conceptually, Hardisty and Wittenberg envisioned the CWFR as the topmost layer of a framework stack (in their white paper). The CWFR in the framework stack is the layer that researchers primarily interact with and contribute to development and standardization.

1.1 Background

The United States National Institutes of Health (NIH) mission is to seek fundamental knowledge about the nature and behavior of living systems, and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability [3]. It conducts and funds focused research on many diseases, including heart, lung, blood, diabetes, digestive, kidney, musculoskeletal, neurological, cancer, and several others.

At the NIH, the National Institute of Neurological Disorders and Stroke (NINDS) funds and supports research programs and activities on many neurological disorders, including Traumatic Brain Injury (TBI) [4] and Parkinson's Diseases (PD) [5]. The NINDS supports investigator-driven research for both TBI and PD, and data is collected from multiple grant-supported research activities.

TBI can result from damage to the brain from external forces and is classified as mild, moderate, or severe. Severe TBI can become life-threatening, and long-term sequelae effect of TBI includes cognitive and physical disability, and post-concussion syndrome, which could lead to chronic traumatic encephalopathy.

Parkinson's Disease results from motor system disorders that result in uncontrollable movements of the body with symptoms that can include tremors, muscle stiffness, bradykinesia (slowing of spontaneous and automatic movement that can make it difficult to perform simple tasks or rapidly perform routine movements),

and postural instability. The exact cause of PD is an area of intense interest; factors like hereditary and a combination of genetic and environmental factors are believed to cause the onset of PD.

A lack of standard data collection methods remains a challenge for disease area research, particularly for TBI. However, efforts are underway by NIH, other US government agencies, and research organizations to address the importance of using common standards, especially for data collection. The TBI research community proposed the use of Common Data Elements during data collection for TBI research [6].

Disease area clinical research is protocol specific and involves the collection of data on specific variables that enable the analyses to prove or disprove research hypotheses. These variables can include data elements or questions that can be aggregated into Case Report Forms (CRFs), which are completed for each patient enrolled in a clinical study. Implementing data elements and electronic CRFs reduces study implementation time and facilitates aggregated data analysis from different sources - including different study sites [7]. Research protocols use commonly defined questions (i.e., variables), CDEs as part of the CRF data collection process [8]. These questions can be associated with specified values and used in a questionnaire or CRF during clinical studies.

The concept of CDEs is independent of a disease area; however, the use of CDEs for a specific disease depends on consensus and adoption by that research community. The advantage of using CDEs is that it standardizes data collection, improves data quality, facilitates data sharing, and provides opportunities for meta-analysis and comparison of results from different studies [9].

The NINDS in collaboration with the disease area research community developed a set of general CDEs that can be used in a variety of clinical studies. The NINDS CDE Project is a collaboration between NINDS and the neurological disorder research community, and provides information on using CDEs, and has also developed CRF templates tailored to research for neurological diseases and disorders [10]. Within the CDE project, a small set of data elements is defined as the Core CDEs and are relevant to all TBI clinical studies and research. For four study types, concussion/mild TBI, acute hospitalized, rehabilitation for moderate/severe TBI, or epidemiological research on TBI, the CRFs include Core and Basic CDEs [11]. Depending on the particulars of the study, Supplemental CDEs may also be used and are available on the template CRFs. Updates to the CDEs are made in collaboration with the research community.

For Parkison's disease CDEs are also defined in terms of Core, Basic, and Supplementary. The CDE Project provides access to PD CDEs, CRF modules, and guidelines. The CRF modules for PD logically organize the CDEs for data collection, while the guidelines provide further information about the CDEs [12].

Here, we examine the application of the concept of CWFR for TBI and PD. We will discuss the use of CDEs as part of a CWFR layer, and provide information on the Biomedical Research Informatics Computing System (BRICS) technical stack, and the Protocol and Form Research Management System (ProFoRMS) used to develop the Federal Interagency Traumatic Brain Injury Research (FITBIR) and Parkinson's Disease Biomarker Program (PDBP) systems. In addition, we explain the role of BRICS components for enabling data objects to be Findable, Accessible, Interoperable, and Reusable (FAIR) [13].

2. CDE-BASED CWFR TECHNICAL STACK

A recurring pattern in both TBI and PD clinical research is hypotheses formulation, data collection, analyses, and dissemination. The FAIR principles state that digital stewardship should promote the discoverability and reuse of digital objects, which include data, metadata, software, and workflows [13]. Furthermore, it posits that data and metadata should be accompanied by Persistent Identifiers, indexed in a searchable resource, retrievable by their identifiers, and use vocabularies that meet domain-relevant community standards. The principles serve as guidelines for improving data discovery and reuse within any infrastructure developed for various research enterprises.

For disease area research, using CDEs during data collection is a way for Digital Objects (DOs) to be reusable as well as to reduce post hoc processing burden during data analyses. Therefore, a CWFR that utilizes CDEs plays an important role in enabling research data to be FAIR.

2.1 Biomedical Research Informatics Computing System

Building systems to support the CDE-based CWFR layer requires an underlying stack comprising technical workflow, execution frameworks, and machinery (see Figure 2 of the white paper). Next, we discuss the implementation of a CWFR layer with an underlying stack of layers that constitute the BRICS, a service-oriented informatics system design that was used in the development of FITBIR and PDBP systems [14].

BRICS comprises infrastructure, business, service, and user interface layers, which provide tools and applications for assisting and creating information packages for maintaining patient confidentiality, metadata association, curation, validation, collection, and subsequent storage within a repository [14]. A variety of tools including the Global Unique Identifier (GUID) client, Translational, Validation, Submission, and Download tools are provided as part of the presentation layer. Individual modules for various components include the Data Dictionary (DD), Account Management, Query Tool, ProFoRMS, Meta Study, Repository Manager, and GUID (Figure 1).

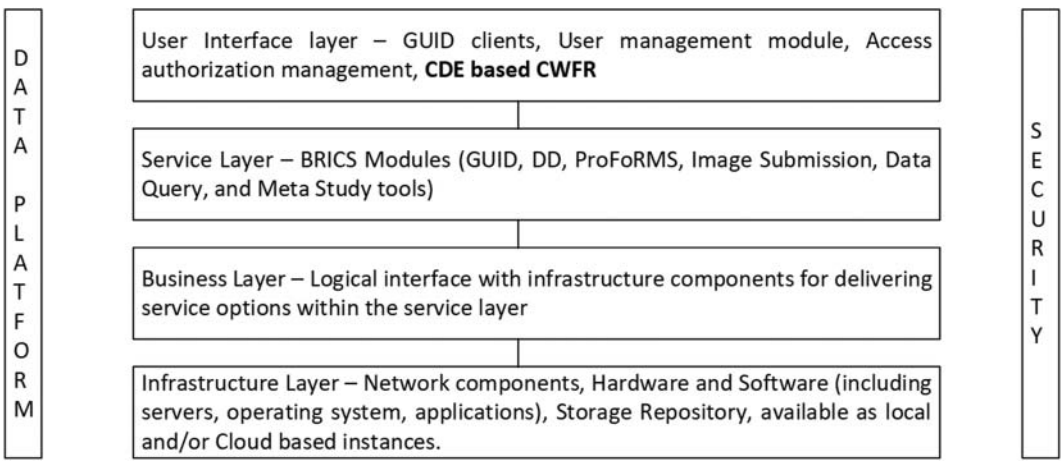

Figure 1. The stack layers for Biomedical Research Informatics Computing System.

The implementation of the BRICS design resulted in FITBIR [15] and PDBP [16] that efficiently utilized the service modules shown in Figure 1, for recurring actions within the TBI and PD clinical research workflows. Both FITBIR and PDBP are supported by the complete stack of layers (shown in Figure 1) that is available at the NIH. Cloud computing can also be used as an infrastructure for deploying various types of workflows for biomedical research [17]. For example, the National Trauma Institute has deployed a cloud-based BRICS instance, the National Trauma Research Repository for supporting new and emerging data sharing needs for the trauma research community [18].

Both TBI and PD researchers can store data within FITBIR and PDBP repositories for long-term preservation and future access [19]. The research community-endorsed CDEs for TBI and PD produce information packages (submission, archive, access) that can be reused for further research. The integration of the CWFR layer with the BRICS results in disease-specific repositories (e.g., FITBIR and PDBP) that are trustworthy, and are consistent with the TRUST principles for digital repositories [20].

2.2 Protocol and Form Research Management System

The BRICS instances for TBI, and PD use the DDs and CDEs with the Protocol and Form Research Management System (ProFoRMS). The main function of ProFoRMS is to optimize the clinical study process by providing functions to researchers for managing protocols, collecting new clinical data using electronic forms, and the submission of collected data for storage in designated repositories [21].

The underlying components of the ProFoRMS that support CDE-based CWFR are illustrated below (Figure 2). The software tools used to execute the ProFoRMS are shown within the Presentation, Application, and Database layers. The ProFoRMS also provides for automatic validation with the data dictionaries of the FITBIR and PDBP. Using the layered approach, introducing enhancements, changes in tools, and upgrades to improve system performance are facilitated.

By using ProFoRMS, a library of files that are comma-separated values (CSV) format is produced and is structured to be consistent with CDE-variable names and values. In addition to FITBIR and PDBP, the ProFoRMS is adaptable for supporting other biomedical research programs.

3. DEVELOPING FAIR DATA OBJECTS (FDOS)

A CDE-based CWFR includes Data Dictionaries, Data Elements, Form Structures (FS), and electronic forms. The CDEs for a disease area can relate to a question on an electronic form and the responses to the questions are made by precisely defined numerical values. The FS serves as the containers for ordered data elements, and the electronic forms are developed using FS as their foundation. A list of Data Elements, FS, and Core CDEs for TBI is available at the FITBIR site [22] and similar resources are available for PD research at the PD site [23].

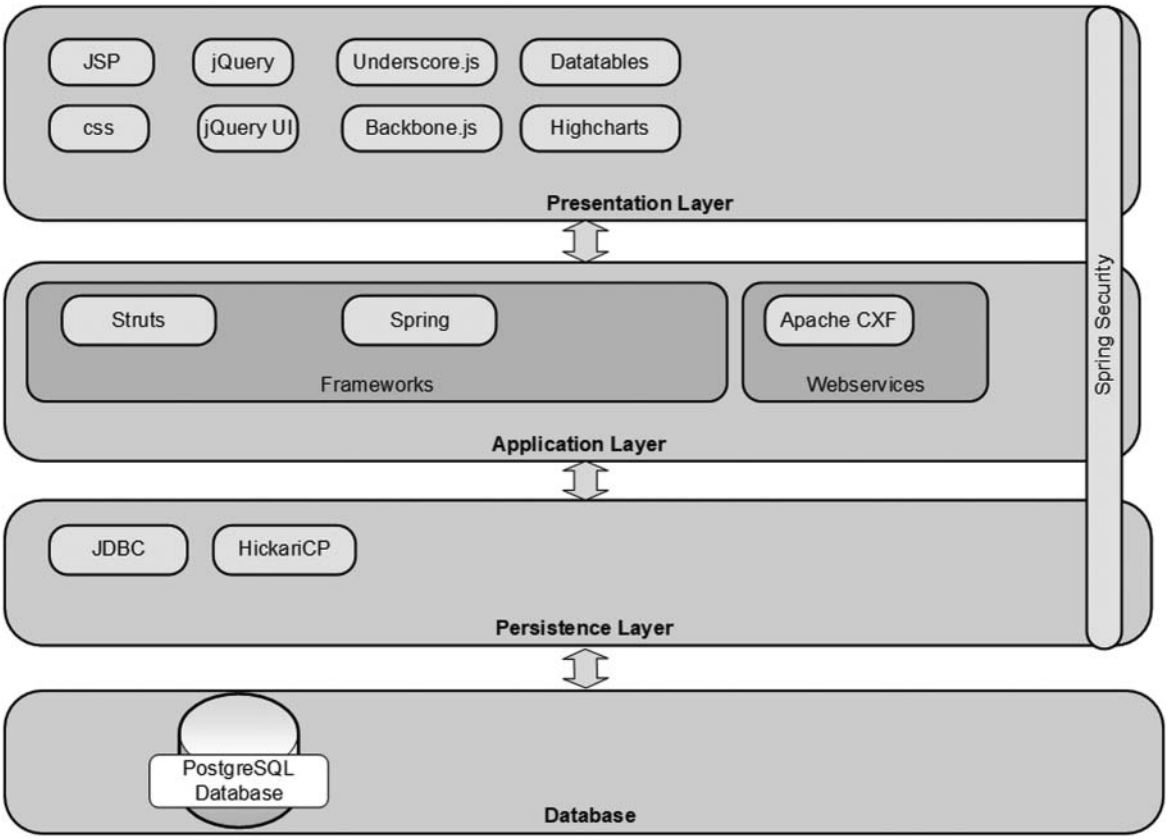

Figure 2. The ProFoRMS components that support a CDE-based CWFR.

The FDOs comprise the essential digital bit sequence, descriptive and technical metadata, and a persistent identifier schema for producing information package(s) that can be stored, preserved, and accessed over time [24]. For TBI and PD research, FDOs are generated for a research study by using the FITBIR and PDBP systems. Protecting personally identifiable information within the workflow process is critical during patient enrollment and throughout the research study. Therefore, before submission to FITBIR or PDBP repository, a canonical step involves the deidentification of patient information. The GUID application program [25] supports the deidentification of data. The program results in computer-generated random alphanumeric code (GUIDs) unique to each patient and stored within the FITBIR and PDBP repositories. The term "Global" in GUID for FITBIR and PDBP systems does not imply the findability of the patient data on the Internet, however, authorized researchers can access the GUIDs to find and link together all submitted data for a single patient who may have been part of multiple clinical studies. For the data set to be an FDO on the web, researchers can organize one or more datasets into a single entity called a 'Study' that serves as a container for the data and metadata with the associated description and methods used for collecting it. Using the DataCite membership available with the FITBIR and PDBP systems, Digital Object Identifiers are assigned for studies to cite in journal articles.

To ensure the proper formatting of files and the formation of FDOs, the Data Mapping and Transformation (DMT) tool [26] is used to translate data received from any user system, for submission to FITBIR and PDBP systems. In addition, the tool saves the translated data into a CSV file that can then be processed by the Validation tool.

The DMT tool is operational on Java-enabled platform such as Windows, UNIX, or Macintosh OS X. Data collected by another generic data collection system (e.g., RedCap) can also be validated with the DMT tool before submission to the FITBIR and PDBP repositories. The FDOs for TBI and PD contain all relevant information needed from the time data is collected, processed, and stored within a specifically designated repository. Additionally, all related resources from a TBI and PD study can be shared as Research Objects (ROs). The ROs are represented as a Meta Study, that includes the results from other studies, and which are aggregated for additional analyses. The ROs have assigned DOIs and are findable and accessible on the Internet.

4. SUMMARY AND CONCLUSION

The NIH supports extensive clinical research programs for many diseases. Clinical research has a recurring workflow pattern that involves hypotheses generation, experimental design, institutional review board approvals, data collection, curation, storage within repositories, analyses, and dissemination of research results. Using CDEs and Data Dictionaries during data collection can result in a CWFR layer that is supported by an informatics system. For disease area research, the underlying layered informatics technical stack for CWFR can comprise the BRICS that process, store, and preserve the data within trusted digital FITBIR and PDBP repositories. The integration of a CWFR layer within an informatics system has resulted in the reuse of software tools and applications in various NIH biomedical research programs. The Traumatic Brain Injury and Parkinson's Disease FDOs contain all the relevant information needed from the time data is first collected to when it is validated and then maintained for future access and use within the designated storage repositories. Using the GUIDs, authorized researchers can access and link together all submitted data for a single patient who may have participated in multiple clinical studies in one or more disease areas. Finally, the use of TBI and PDBP data dictionaries results in DOs that are findable, accessible, and reusable for research.

ACKNOWLEDGMENTS

The authors thank Alicia A. Livinski, NIH Library, for manuscript editing assistance and Tsega Gebremichael, Sapient Government Services, for illustrating the ProFoRMS figure.

AUTHOR CONTRIBUTIONS

V. Navale (Vivek.Navale@nih.gov) conceptualized and developed the manuscript. He was responsible for the writing of the original draft, reviewing, revising the manuscript, and addressing reviewers' comments. M. McAluiffe (mcmatt@exchange.nih.gov) provided the details for the Protocol and Form Research Management System and also reviewed and edited the manuscript.

REFERENCES

[1] Navale, V., von Kaeppler, D., McAuliffe, M.: An overview of biomedical platforms for managing research data. Journal of Data, Information and Management 3, 21–27 (2021)
[2] Hardisty, A., Wittenburg, P. (eds.): Canonical Workflow Framework for Research (CWFR)—position paper—version 2, December 2020. Working paper. Available at: https://osf.io/9e3vc/. Accessed 7 December 2021
[3] National Institutes of Health (NIH). Available at: https://www.nih.gov/. Accessed 7 December 2021
[4] Focus on traumatic brain injury research. Available at: https://www.ninds.nih.gov/Current-Research/Focus-Disorders/Traumatic-Brain-Injury. Accessed 7 December 2021
[5] Parkinson's disease biomarkers program. Available at: https://pdbp.ninds.nih.gov/. Accessed 7 December 2021
[6] Thompson, H.J., Vavilala, M.S., Rivara, F.P.: Common data elements and federal interagency traumatic brain injury research informatics system for TBI research. Annual Review of Nursing Research 33(1), 1–11 (2015)
[7] Navale, V., et al.: Standardized informatics computing platform for advancing biomedical discovery through data sharing. bioRxiv preprint bioRxiv: 259465 (2018)
[8] Common Data Element (CDE)—Clinfowiki. Available at: https://clinfowiki.org/wiki/index.php/Common_Data_Element_(CDE). Accessed 7 December 2021
[9] Sheehan, J., et al.: Improving the value of clinical research through the use of Common Data Elements. Clinical Trials: Journal of the Society for Clinical Trials 13(6), 671–676 (2016)
[10] NINDS Common Data Elements. Available at: https://commondataelements.ninds.nih.gov/. Accessed 7 December 2021
[11] Traumatic brain injury. Available at: https://www.commondataelements.ninds.nih.gov/Traumatic%20Brain%20Injury. Accessed 7 December 2021
[12] Parkinson's disease. Available at: https://commondataelements.ninds.nih.gov/Parkinson%27s%20Disease#pane-158. Accessed 7 December 2021
[13] Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3, 160018 (2016)
[14] Navale, V., et al.: Development of an informatics system for accelerating biomedical research. F1000Research 8, 1430 (2019)
[15] Federal interagency traumatic brain injury research. Available at: https://fitbir.nih.gov/. Accessed 7 December 2021
[16] The NINDS PDBP data management resource. Available at: https://pdbp.ninds.nih.gov/data-management-resource. Accessed 7 December 2021
[17] Navale, V., Bourne, P.E.: Cloud computing applications for biomedical science: A perspective. PLoS Computational Biology 14, e1006144 (2018)
[18] Price, M.A., et al.: Launch of the National Trauma Research Repository coincides with new data-sharing requirements. Trauma Surgery & Acute Care Open 3, e000193 (2018)

[19] Navale, V., McAuliffe, M.: Long-term preservation of biomedical research data. F1000Research 7, 1353 (2018)
[20] Lin, D., et al.: The TRUST principles for digital repositories. Scientific Data 7, 144 (2020)
[21] Introducing BRICS. Available at: https://brics.cit.nih.gov/intro. Accessed 7 December 2021
[22] FITBIR data dictionary. Available at: https://fitbir.nih.gov/content/data-dictionary#data-elements. Accessed 7 December 2021
[23] Parkinson's disease. Available at: https://commondataelements.ninds.nih.gov/Parkinson%27s%20Disease. Accessed 7 December 2021
[24] De Smedt, K., Koureas, D., Wittenburg, P.: FAIR digital objects for science: From data pieces to actionable knowledge units. Publications 8(2), Article No. 21 (2020)
[25] Johnson, S.B., et al.: Using global unique identifiers to link autism collections. Journal of American Medical Informatics Association (JAMIA) 17, 689–695 (2010)
[26] BRICS Data Mapping and Transformation tool. Available at: https://brics.cit.nih.gov/sites/default/files/bricsPdf/chapter_12_-_data_mapping_and_transformation_tool.pdf. Accessed 7 December 2021

AUTHOR BIOGRAPHY

Dr. **Vivek Navale** joined the National Institutes of Health (NIH), United States of America (USA) in 2014. He serves as a senior manager for the NIH Center for Information Technology and has been involved in the development of informatics systems for biomedical research programs. Before NIH, he was the Chief of Information Technology Services for the USA National Archives and Records Administration (NARA), for 14 years where he was responsible for digital infrastructure design, development, and management of data repositories. During 1990–2000, he worked as a Principal Scientist with the Raytheon Technical Services Company and at the NASA Goddard Space Flight Center. His contributions included sensor design, development, and integration with the space-bound instruments for NASA's Cassini Mission to Saturn and its moon Titan. Over the past three decades, Dr. Navale has presented, chaired several national and international conferences. He also serves as a scientific reviewer for many journals and NIH-sponsored research programs and has published research papers in a variety of scientific journals. He was the recipient of multiple awards—the NIH Director, the NARA Archivist, and the NASA Cassini Mission team awards. He earned his doctorate in Chemistry from George Washington University, Washington, DC, USA.
ORCID: 0000-0002-7110-8946

Dr. **McAuliffe** has been at NIH since 1998 and is currently the Chief of the Scientific Application Services (SAS) section in the Office of Scientific Computing Services (OSCS). He provides computational and engineering expertise to a variety of clinical and biomedical informatics activities at NIH and is committed to supporting data sharing and making data FAIR (Findable, Accessible, Interoperable, and Reusable). He leads the development of the Biomedical Research Informatics Computing System (BRICS) (http://brics.cit.nih.gov/) which is a comprehensive data informatics system designed to efficiently collect, validate, harmonize, and analyze research datasets. In addition, Dr. McAuliffe strives to advance and empower scientific imaging research in the NIH intramural program, to this end, SAS has created and continues the development of the successful Medical Image Processing Analysis and Visualization application (MIPAV: http://mipav.cit.nih.gov). Dr. McAuliffe earned his Bachelor of Science in Electrical Engineering degree from the University of Detroit, Master's in Software Engineering from Johns Hopkins, and his Ph.D. in Biomedical Engineering from the University of North Carolina at Chapel Hill, NC. His current research interests include medical informatics and biomedical imaging specializing in segmentation, quantification, and image fusion.
ORCID: 0000-0002-2409-5126